15400 Calhoun Drive, Suite 400
Rockville, Maryland, 20855
(301) 294-5200
http://www.i-a-i.com

Intelligent Automation Incorporated

Enhancements for a Dynamic Data Warehousing and Mining System for Large-Scale Human Social Cultural Behavioral (HSBC) Data

Final Report

Reporting Period: March 22, 2016 – September 18, 2016

Contract No. N00014-16-P-3014

Sponsored by
ONR, Arlington VA
COTR/TPOC: Dr. Rebecca Goolsby

Prepared by
Onur Savas, Ph.D.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*
26/09/2016

2. REPORT TYPE
Final Report

3. DATES COVERED *(From - To)*
March 22, 2016 – September 18, 2016

4. TITLE AND SUBTITLE
Enhancements for a Dynamic Data Warehousing and Mining System for Large-Scale Human Social Cultural Behavioral (HSBC) Data

5a. CONTRACT NUMBER
N00014-16-P-3014

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Onur Savas

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Intelligent Automation, Inc.
15400 Calhoun Drive, Suite 190
Rockville, MD 20855

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research (BD253)
875 N. Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Report contains color

14. ABSTRACT

In this project, to continue our development of a system that can dynamically collect and warehouse unfiltered textual communication data and make this data available to support HSCB modeling, decision making, and course of action development, we have (i) developed algorithms and an end-to-end computational framework to support graph analysis and visualization, (ii) developed graph querying capabilities such as top K and n-hop neighborhood, (iii) developed automated YouTube video metadata collection and analysis of YouTube data, and (iv) developed automated VK data collection and analysis of VK data. The features are released as part of Scraawl 2.0.

15. SUBJECT TERMS
social media, HSCB, data warehousing, large-scale analysis, OSINT

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Rebecca Goolsby
Unclassified	Unclassified	Unclassified	SAR	15	19b. TELEPHONE NUMBER *(Include area code)* (703) 588-0558

Table of Contents

1 Executive Summary

In this project, to continue our development of a system that can dynamically collect and warehouse unfiltered textual communication data and make this data available to support HSCB modeling, decision making, and course of action development, we have accomplished the following.

First, we have developed algorithms and an end-to-end computational framework to ingest Twitter data, convert it into Twitter interaction and/or friendship graphs, and store in a graph database, namely OrientDB. We have also enhanced Scraawl's current graph visualization for scalability and usability.

Second, we have developed graph querying capabilities and implemented various UIs to visualize the outputs of these queries. In particular, we have developed top K user and/or hashtag query, and n-hop neighborhood queries. Various attributes of the nodes and edges can also be interactively visualized.

Third, we have developed automated YouTube video metadata collection capabilities. In particular, we have developed two APIs for continuous (Streaming) and recent YouTube data searches, and designed a user-friendly UI for these searches using YouTube Data API v3. Algorithms to perform basic statistics computation, NER, and map visualization have also been designed and implemented.

Fourth, we have developed automated VK data collection capabilities. Algorithms to perform basic statistics computation, NER, and map visualization have also been designed and implemented.

Finally, we released Scraawl 2.0, which incorporates these features.

2 Technical Work

2.1 Store Twitter Interaction Graphs in OrientDB

The first step in developing the querying capability is to store a graph efficiently. A graph database, for all practical purposes, can be represented as the graph itself. In its simplest form, we consider the *Twitter interaction graph* modeled as an undirected graph $G = (V, E)$, where V is the set of nodes (vertices) and E is the set of edges. For a collection of tweets $\{\tau(\theta)|\ \theta\ \in \mathbb{Z}^+\}$, where each tweet $\tau(.)$ can be uniquely identified by its unique *tweet ID* $\theta\ \in \mathbb{Z}^+$, let $\tau(\theta)$ be tweeted by user $u_{\tau(\theta)}$ and let $u_{\tau(\theta)}$ have retweeted, mentioned, or replied to K_θ users $\mathcal{I}\big(\tau(\theta)\big) = \{v^1_{\tau(\theta)}, v^2_{\tau(\theta)}, \dots, v^{K_\theta}_{\tau(\theta)}\}$. Of course, if no retweets, mentions, or replies are present, then $\mathcal{I}\big(\tau(\theta)\big) = \emptyset$. We can then unambiguously specify the Twitter Interaction Graph G with $V = \big\{u_{\tau(\theta)} \cup \mathcal{I}\big(\tau(\theta)\big)\big|\ \theta\ \in \mathbb{Z}^+\}$ and $E = \big\{u_{\tau(\theta)} \times \mathcal{I}\big(\tau(\theta)\big)\big|\ \theta\ \in \mathbb{Z}^+\} = \quad \{(u_{\tau(\theta)}, v^1_{\tau(\theta)}), (u_{\tau(\theta)}, v^2_{\tau(\theta)}), \dots, (u_{\tau(\theta)}, v^K_{\tau(\theta)})|\theta\ \in \mathbb{Z}^+\}$. One can enhance G by adding other types of interactions, e.g., by adding an edge if a user uses a hashtag in his/her tweet. In this particular case, the vertex set will have hashtags as well. Note that this definition can trivially be extended for *friendship graphs*, where the interactions between users/nodes are whether they are friends and/or followers.

To implement the graph modeling capabilities above, we used an open source graph database, namely OrientDB (http://www.orientdb.com), and an SQL-like graph querying language. OrientDB provides an NoSQL engine that stores and queries graphs via both (i) a graph database API and document API, and (ii) supports schema-less, schema-full, and schema-mixed modes. Our initial experimentation with OrientDB, an open source graph database, and its SQL-like graph querying language yielded promising results. In particular, we inserted a Twitter Interaction Graph with $|V| > 34k$ and $|E| > 421k$ in less than 6 minutes. In addition, we ordered nodes of this graph by degree in less than 2 seconds. A basic graph service that performs basic graph database management operations, e.g., insert, delete, update, is also implemented.

We have also designed and implemented the *k*-hop neighbor queries. In brief and informally, starting from a node *v*, *k*-hop queries finds the neighbors of *v*, neighbors of neighbors of *v* (a.k.a. second degree neighbors), and so on until *k*-hop neighbors. In particular, for a querying function *Q(.),* a "2-hop" neighbors query is implemented as follows.

$$Q(v) = \big\{(v, v') \cup \big(v', v''\big)\big|(v, v') \in E\ and\ (v', v'') \in E\ for\ \forall\ v', v'' \in V\big\}.$$

This can be generalized to *k*-hop by including neighbors of neighbors of neighbors up *k*-hop.

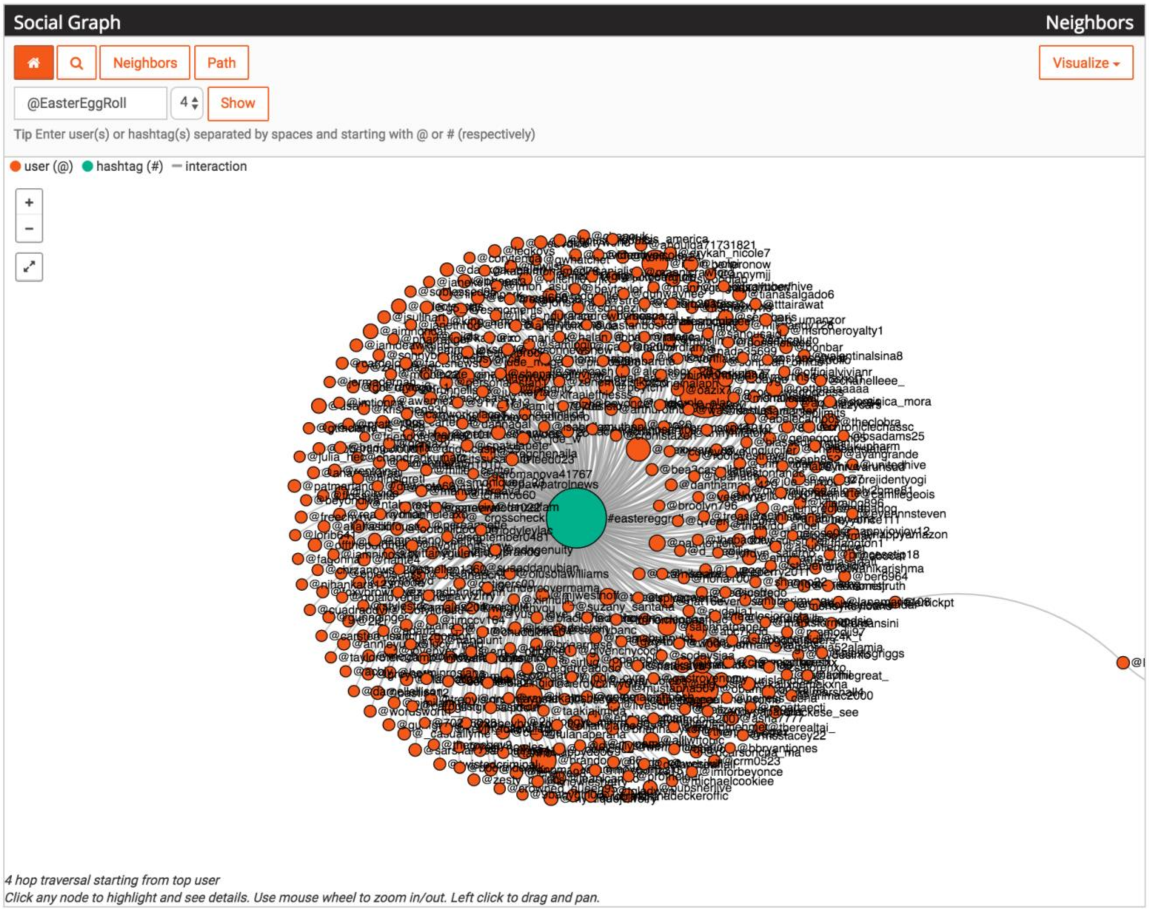
This querying service has also been implemented using a Groovy/Grails framework with functions written in Java.

2.2 Design and Implement Visualization and UI

To visualize the graph returned by the queries, we have further developed IAI's graph visualization capabilities. We have also incorporated a UI to call the *k*-hop query service. These capabilities are all incorporated to Scraawl.

Figure 1 is a visualization for 4-hop neighborhood query starting from #EasterEggRoll. The Twitter interaction graph has been created from a collection of tweets that have been collected between 28 Mar, 2016 03:00 AM and 29 Mar, 2016 02:59 using the keyword #EasterEggRoll. Overall, 25063 tweets were collected. The graph has 16171 nodes (vertices) and 48594 edges.

The "Neighbors" button to the upper left corner call the k-hop querying service with (i) starting node (in this case #EasterEggRoll), and (ii) the *k*-parameter (in this case *k* = 4).



@EasterEggRoll
4

Tip Enter user(s) or hashtag(s) separated by spaces and starting with @ or # (respectively)

4 hop traversal starting from top user
Click any node to highlight and see details. Use mouse wheel to zoom in/out. Left click to drag and pan.

Figure 1: 4-hop neighborhood query starting from "#eastereggroll".

2.3 Top K User and Hashtag Subgraph Querying and Visualization

2.3.1 Top K Query Implementation

As reported in Progress Report No. 1, we consider the Twitter interaction graph modeled as an undirected graph $G = (V, E)$, where V is the set of nodes (vertices) and E is the set of edges. For a collection of tweets $\{\tau(\theta) | \theta \in \mathbb{Z}^+\}$, where each tweet $\tau(.)$ can be uniquely identified by its unique *tweet ID* $\theta \in \mathbb{Z}^+$, let $\tau(\theta)$ be tweeted by user $u_{\tau(\theta)}$ and let $u_{\tau(\theta)}$ have retweeted, mentioned, or replied to K_θ users $\mathcal{J}(\tau(\theta)) = \{v^1_{\tau(\theta)}, v^2_{\tau(\theta)}, \ldots, v^{K_\theta}_{\tau(\theta)}\}$. Of course, if no retweets, mentions, or replies are present, then $\mathcal{J}(\tau(\theta)) = \emptyset$. We can then unambiguously specify the Twitter Interaction Graph G with $V = \{u_{\tau(\theta)} \cup \mathcal{J}(\tau(\theta)) | \theta \in \mathbb{Z}^+\}$ and $E = \{u_{\tau(\theta)} \times \mathcal{J}(\tau(\theta)) | \theta \in \mathbb{Z}^+\} = \{(u_{\tau(\theta)}, v^1_{\tau(\theta)}), (u_{\tau(\theta)}, v^2_{\tau(\theta)}), \ldots, (u_{\tau(\theta)}, v^K_{\tau(\theta)}) | \theta \in \mathbb{Z}^+\}$.

Our interest lies in the querying of this graph. We first define a generic query operator over the graph that will return a subset of the graph, i.e., $Q(.): G \mapsto G$. In this reporting period, we have designed and implemented top[1] K neighborhood queries. Formally, we define the top K neighborhood query as

$$Q\left(\bigcup_i v_i\right) = G(V', E')\ where\ V' = \left(\bigcup_i v_i \cup N(v_i)\right) and\ s.t.\ E' = \bigcup_{i,j} (v_i, v_j(i)),$$
$$where\ v_j(i) \in N(v_i)$$

for a set of "top" $\bigcup_i v_i$ vertices and $N(v_i)$ denotes the immediate neighborhood of v_i. Recall from Report 1 that we have designed and implemented a graph querying system using the graph database OrientDB. We have added the above querying capability implemented using OrientDB and made it available through Scraawl.

In particular, we have implemented top K neighborhood queries for the following top K.

1. **Top K Users:** The top tweeting users.

2. **Top K Hashtags:** The top tweeted hashtags.

3. **Top Connected Users:** Top users ranked by degree in the Social Graph.

4. **Top Connected Hashtags**: Top hashtags ranked by degree in the Social Graph.

We have also enhanced the visualization for top K queries. We have improved the visualization by (i) automatically adjusting zooming and translating so that the graph fits into the visible screen, (ii) incorporating web workers, i.e., background computation threads, so that graph layout is computed without interfering with user's interaction with the page, and (iii) adjusted the parameters of the graph layout so that there is less

[1] As defined according to context as shown in Section 1.1.2.

cluttering. Figure 2 is a representative visualization for a top 10 query as integrated into Scraawl.

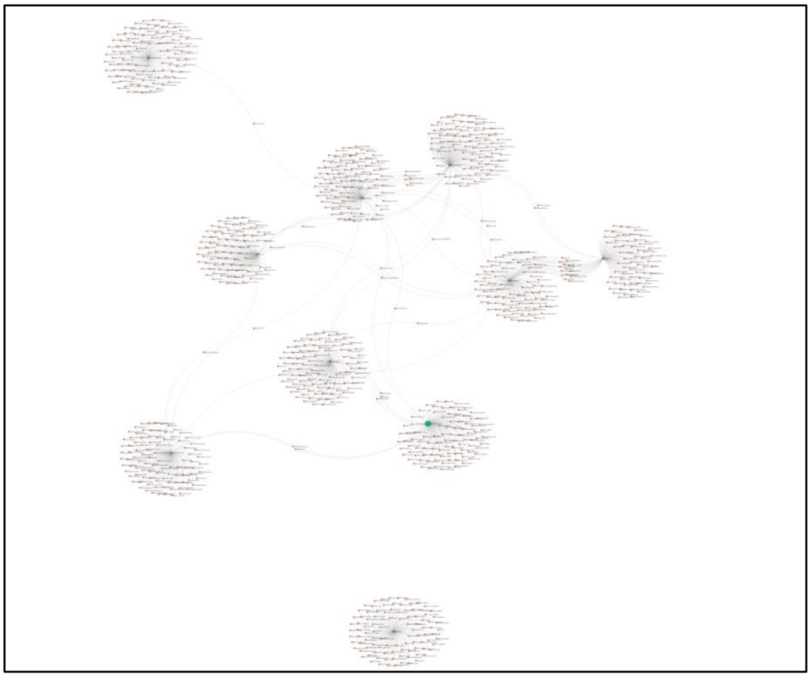


Figure 2: Representative visualization of Top 10 users.

2.3.2 Graph Visualization for Friendship Graph

We have also designed and implemented a service that retrieves the friends and followers of a given user and visualizes it in Scraawl. Figure 3 is a representative graph visualization of the friendship graph of two users. The “queried users” are depicted as orange, the common friends/followers are shown as green, and the rest of the users are drawn in gray. Unlike the Social graph, the friendship graph is directed hence the direction of links are represented in the visualization. The relationship “being a friend” and “being a follower” are distinguished by using orange and gray colors.

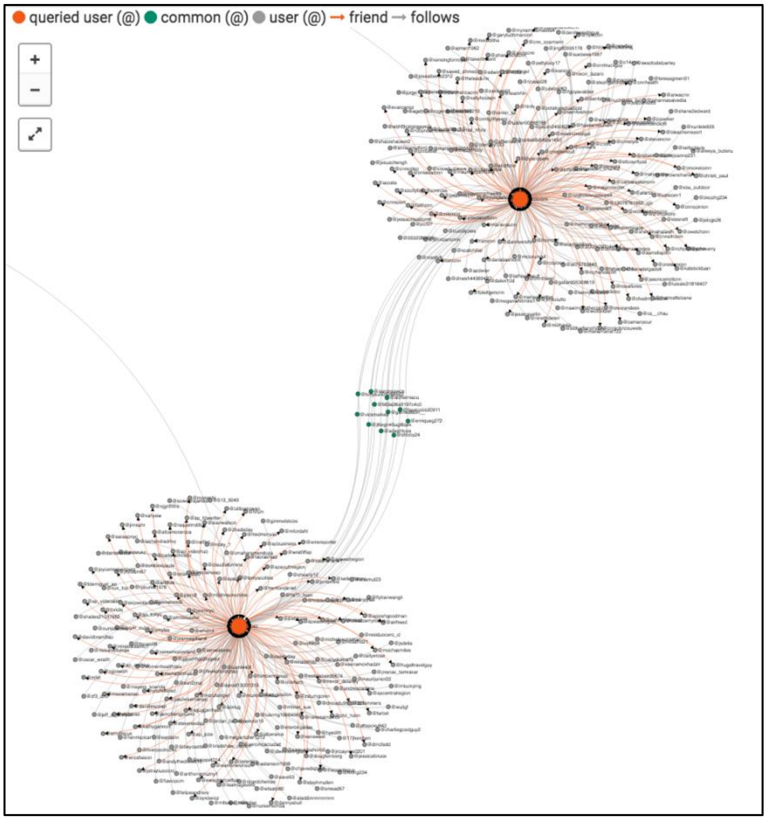

● queried user (@) ● common (@) ● user (@) → friend → follows
+
−

Figure 3: Representative Friendship Graph.

2.4 YouTube Data Collection and Analytics

2.4.1 YouTube Data Collection

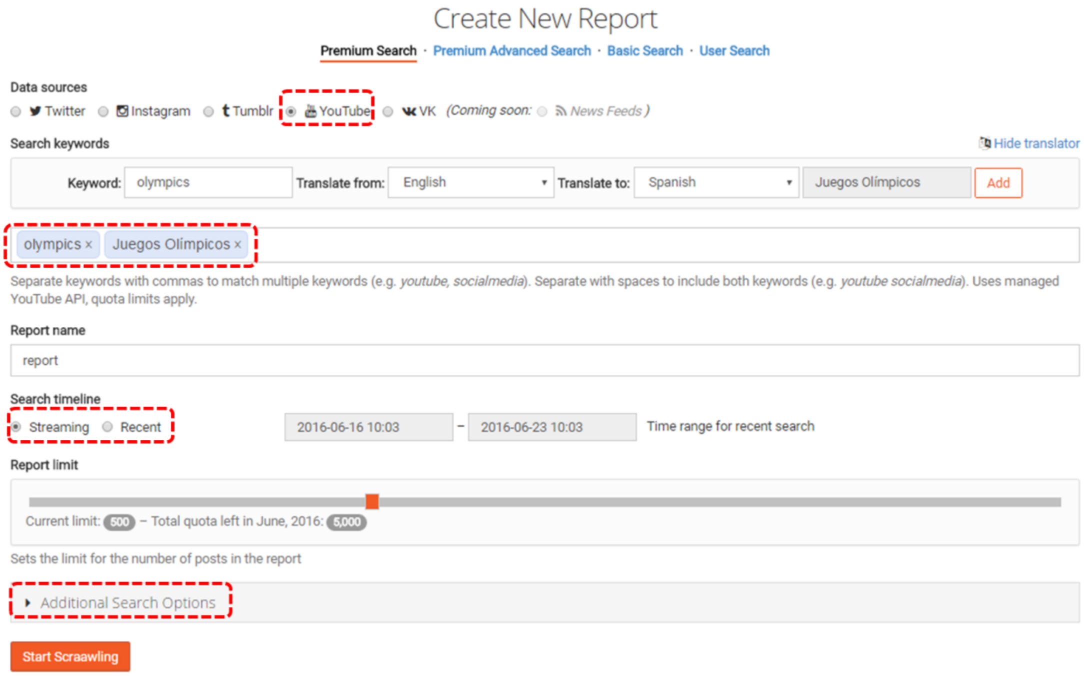
The first API allows for streaming searches, i.e., adds a query for continuous YouTube post collection. The second API allows searching on recent YouTube videos, and the resulting data will be saved in the configured database. Both APIs make use of the (i) `Search:list` Data API functionality, which returns a collection of search results that match the query parameters specified in the API request, and (ii) `Videos: list` Data API functionality, which returns a list of videos that match the API request parameters. Some of the major parameters used in the Scraawl streaming and recent API searches is shown in Table 1.

Table 1: Scraawl API Major YouTube Data Collection Parameters.

Parameter	Explanation
query.q	The \`q\` parameter specifies the query term to search for. Two words separated by spaces can be treated as AND. Your request can also use the Boolean NOT (-) and OR (|) operators to exclude videos or to find videos that are associated with one of several search terms.
query.channelId	The \`channelId\` parameter indicates that the API response

	should only contain resources created by the channel.
query.location & query.locationRadius	The `location` parameter, in conjunction with the `locationRadius` parameter, defines a circular geographic area and also restricts a search to videos that specify, in their metadata, a geographic location that falls within that area.
query.publishedBefore	The `publishedBefore` parameter indicates that the API response should only contain resources created before the specified time.
query.publishedAfter	The `publishedBefore` parameter indicates that the API response should only contain resources created before the specified time.

We have also developed a UI to use the above APIs seamlessly. A representative UI is shown in Figure 4. The UI has the same look and feel with other social media searches, and can be accessed from "Create New Report" view under Scraawl. The UI allows to specify keywords with each box "AND"ed, and the translation capability using Google translate is integrated. The user can also choose between "Streaming" and "Recent" searches, which in turn calls one of the above APIs explained in Section 1.1.1. When "Recent" search is selected, the user has the option to select a time range. When "Streaming" search is selected, data collection will continue until a pre-specified time-out or the data collection limit is reached. Similar to other data feeds, we also allow the user to draw circular bubbles on the world map to restrict their searches to certain region(s) under "Additional Search Options."



Create New Report

Premium Search · Premium Advanced Search · Basic Search · User Search

Data sources

○ Twitter ○ Instagram ○ Tumblr ● YouTube ○ VK (Coming soon: ○ News Feeds)

Search keywords

Hide translator

Keyword: olympics Translate from: English Translate to: Spanish Juegos Olímpicos Add

olympics × Juegos Olímpicos ×

Separate keywords with commas to match multiple keywords (e.g. *youtube, socialmedia*). Separate with spaces to include both keywords (e.g. *youtube socialmedia*). Uses managed YouTube API, quota limits apply.

Report name

report

Search timeline

● Streaming ○ Recent 2016-06-16 10:03 – 2016-06-23 10:03 Time range for recent search

Report limit

Current limit: 500 – Total quota left in June, 2016: 5,000

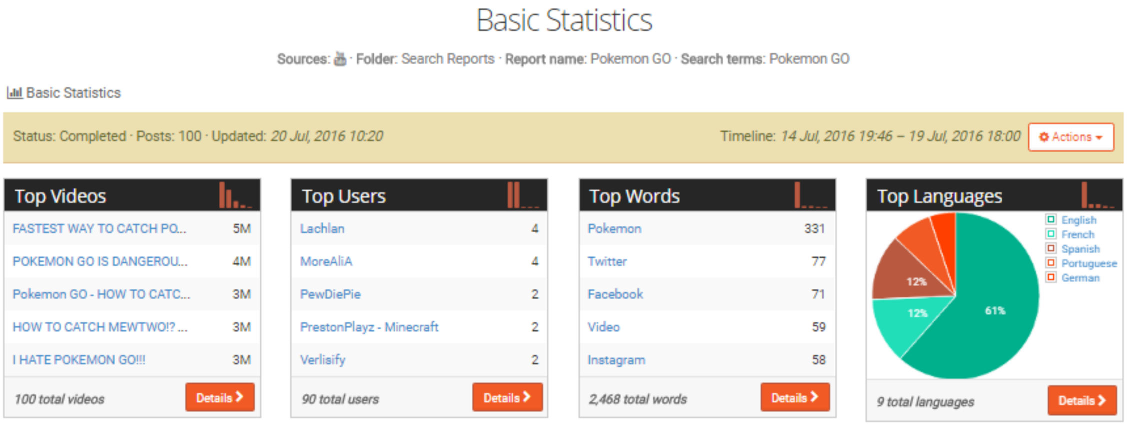
Sets the limit for the number of posts in the report

Additional Search Options

Figure 4: UI for YouTube Searches.

2.4.2 YouTube Analytics

We have developed a fast and reliable computation of Top K statistics for YouTube videos. In particular, we compute Top Videos, Top Users, Top Words, and Top Languages. In all cases, “Top” refers to higher count. Similar to other data sources, we have also implemented a UI to show and interact with these statistics. Figure 5 shows a representative view of the UI. Top Videos, Top Users, and Top Words are shown as lists while Top Languages are shown as a pie chart. Every Top K Statistics also include a drill-down menu, where users can access using the “Details” button. In the drill-down menu, users can see Top 50 statistics and other relevant metrics. Also, users can filter to include or exclude the relevant posts.



Basic Statistics

Sources: · Folder: Search Reports · Report name: Pokemon GO · Search terms: Pokemon GO

Basic Statistics

Status: Completed · Posts: 100 · Updated: *20 Jul, 2016 10:20*

Timeline: *14 Jul, 2016 19:46 – 19 Jul, 2016 18:00*

Top Videos	
FASTEST WAY TO CATCH PO...	5M
POKEMON GO IS DANGEROU...	4M
Pokemon GO - HOW TO CATC...	3M
HOW TO CATCH MEWTWO!? ...	3M
I HATE POKEMON GO!!!	3M
100 total videos	Details

Top Users	
Lachlan	4
MoreAliA	4
PewDiePie	2
PrestonPlayz - Minecraft	2
Verlisify	2
90 total users	Details

Top Words	
Pokemon	331
Twitter	77
Facebook	71
Video	59
Instagram	58
2,468 total words	Details

Top Languages

9 total languages Details

Figure 5: Representative Top K Statistics for YouTube Videos.

Similar to the other data feeds, we have developed a timeline view that shows the number of posts. The timeline view is interactive and the selected portion can be filtered using “Filter By Range” as shown in Figure 6.

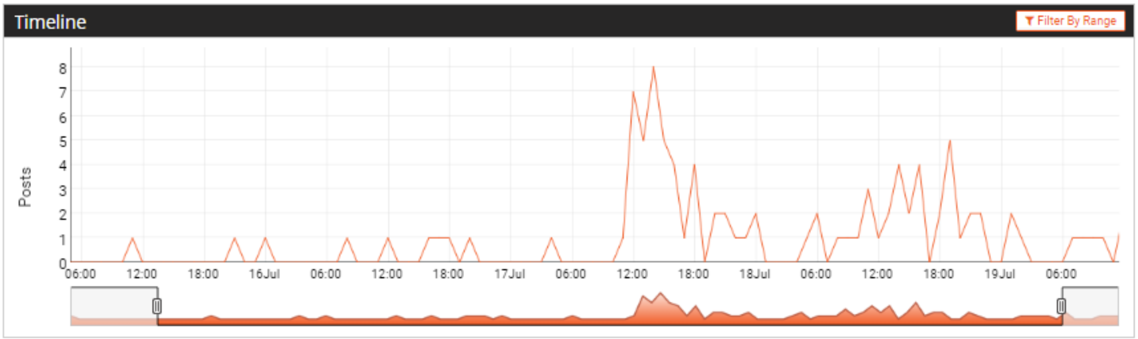


Figure 6: Representative Timeline View.

We show the snapshots of the Top 50 videos in a grid under “Media Gallery” view. The video snapshots are clickable and each click directs the user to a page that has statistics about video. A representative media gallery view is shown in Figure 7.

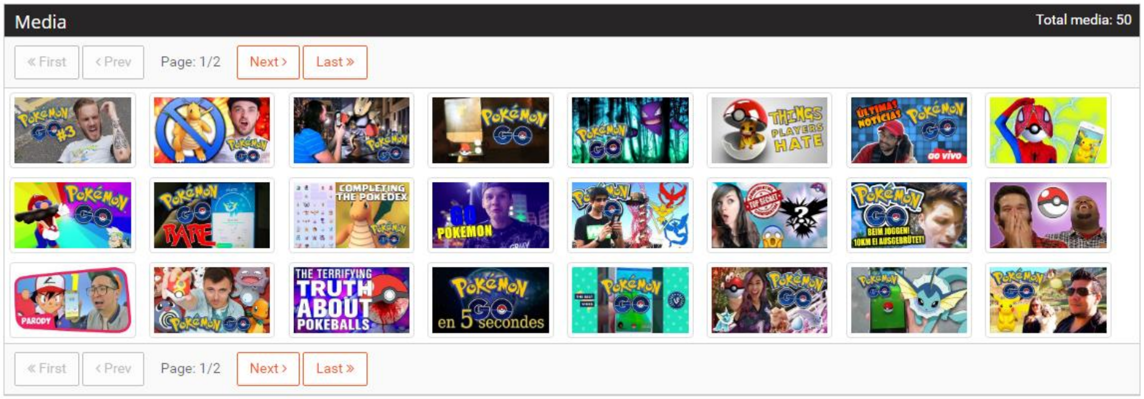
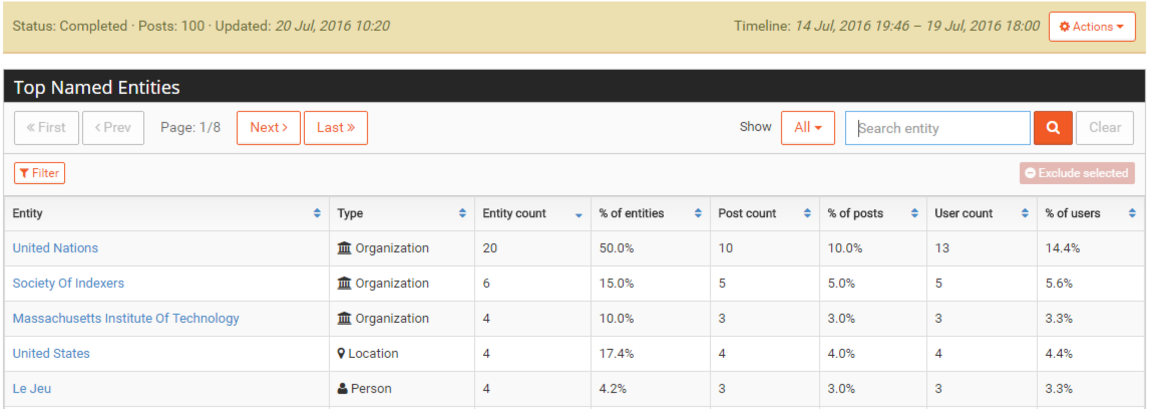


Figure 7: Representative Media Gallery View.

We perform Scraawl's NER algorithm to the text associated with YouTube post, which classifies the named entities into organizations, locations, and persons. Scraawl's NER also performs abbreviation extensions, e.g., UN is mapped to "United Nations." A representative NER view along with other statistics is shown in Figure 8.



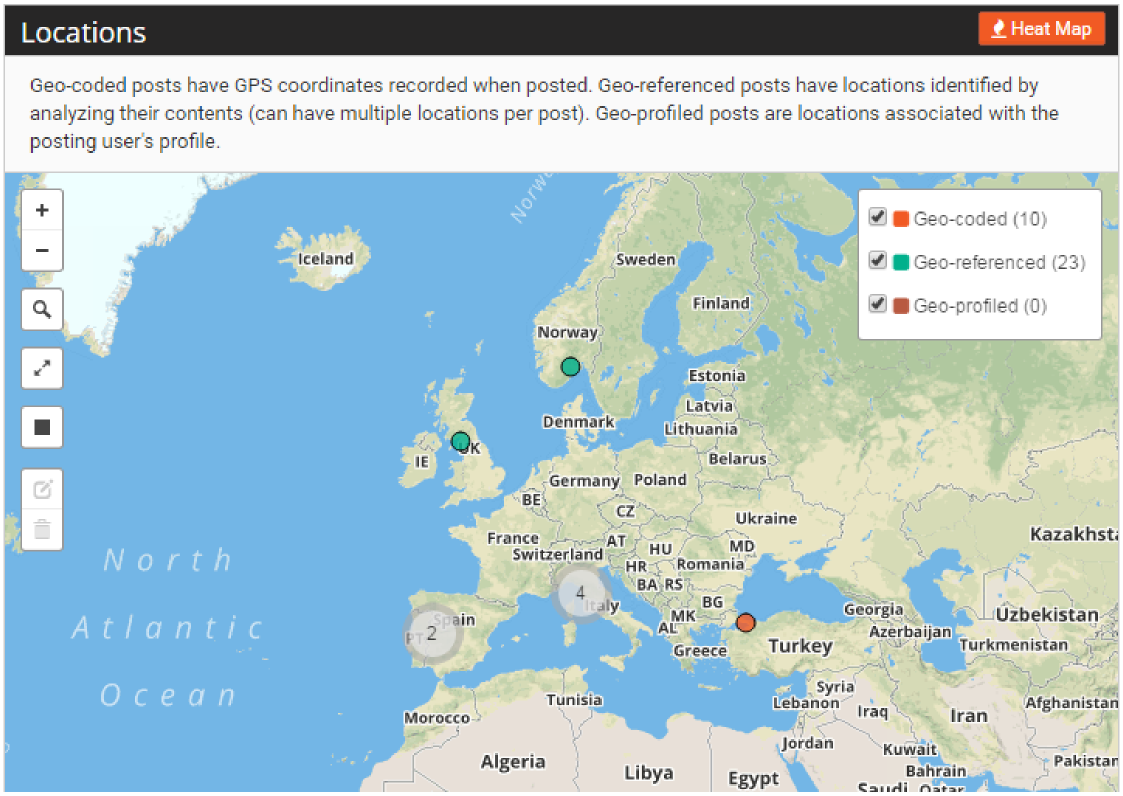
Status: Completed · Posts: 100 · Updated: *20 Jul, 2016 10:20*

Timeline: *14 Jul, 2016 19:46 – 19 Jul, 2016 18:00*

Entity	Type	Entity count	% of entities	Post count	% of posts	User count	% of users
United Nations	Organization	20	50.0%	10	10.0%	13	14.4%
Society Of Indexers	Organization	6	15.0%	5	5.0%	5	5.6%
Massachusetts Institute Of Technology	Organization	4	10.0%	3	3.0%	3	3.3%
United States	Location	4	17.4%	4	4.0%	4	4.4%
Le Jeu	Person	4	4.2%	3	3.0%	3	3.3%

Figure 8: Representative NER View.

We use Scraawl's Location Map and Heat Map view to display geolocation information about YouTube videos. In particular, we show geo-coded and geo-referenced YouTube posts. Geo-coded posts are those which have GPS information on them while geo-referenced posts are those which mention location information. A representative Map view is shown in Figure 9.



Locations

Heat Map

Geo-coded posts have GPS coordinates recorded when posted. Geo-referenced posts have locations identified by analyzing their contents (can have multiple locations per post). Geo-profiled posts are locations associated with the posting user's profile.

☑ 🟧 Geo-coded (10)

☑ 🟩 Geo-referenced (23)

☑ 🟥 Geo-profiled (0)

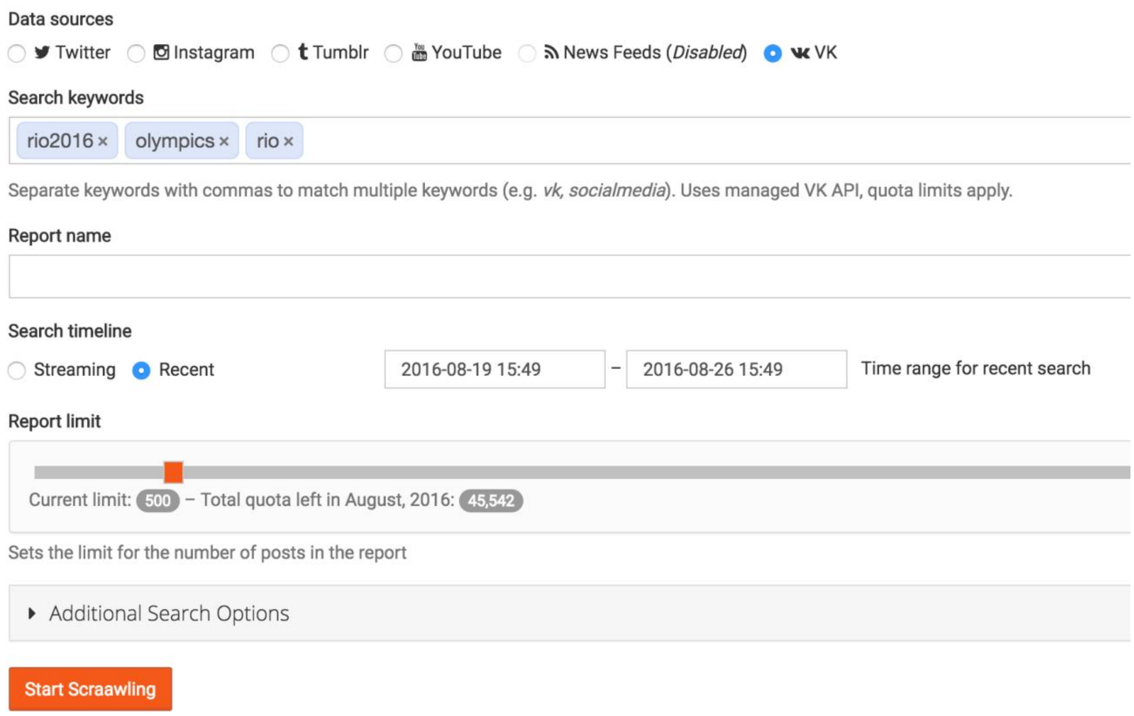
Figure 9: Representative Map View.

2.5 VK Data Collection and Analytics

2.5.1 VK Data Collection

VK is the largest European online social networking service. It is available in several languages, but is especially popular among Russian-speaking users. Similar to other social networks, VK allows users to message each other publicly or privately, to create groups, public pages and events, share and tag images, audio and video, and to play browser-based games.

In this reporting period, we have matured VK crawling capabilities. In particular, we have developed a capability to retrieve VK posts based on keywords. Current search grammar includes combining a set of keywords by AND'ing or OR'ing them. The search can be performed as streaming or a 1-week historical, and can be combined with a geo-location search. The search is integrated as part of Scraawl, and the UI is shown in Figure 10. Similar to other social network searches, the keywords or phrases can be entered separately, a report name can be given to the search, and either a streaming or a historical search can be chosen. In addition, a map is interactively used (not shown) to restrict the search to a geospatial region.



Data sources

○ Twitter ○ Instagram ○ Tumblr ○ YouTube ○ News Feeds (*Disabled*) ● VK

Search keywords

rio2016 × olympics × rio ×

Separate keywords with commas to match multiple keywords (e.g. *vk, socialmedia*). Uses managed VK API, quota limits apply.

Report name

Search timeline

○ Streaming ● Recent 2016-08-19 15:49 – 2016-08-26 15:49 Time range for recent search

Report limit

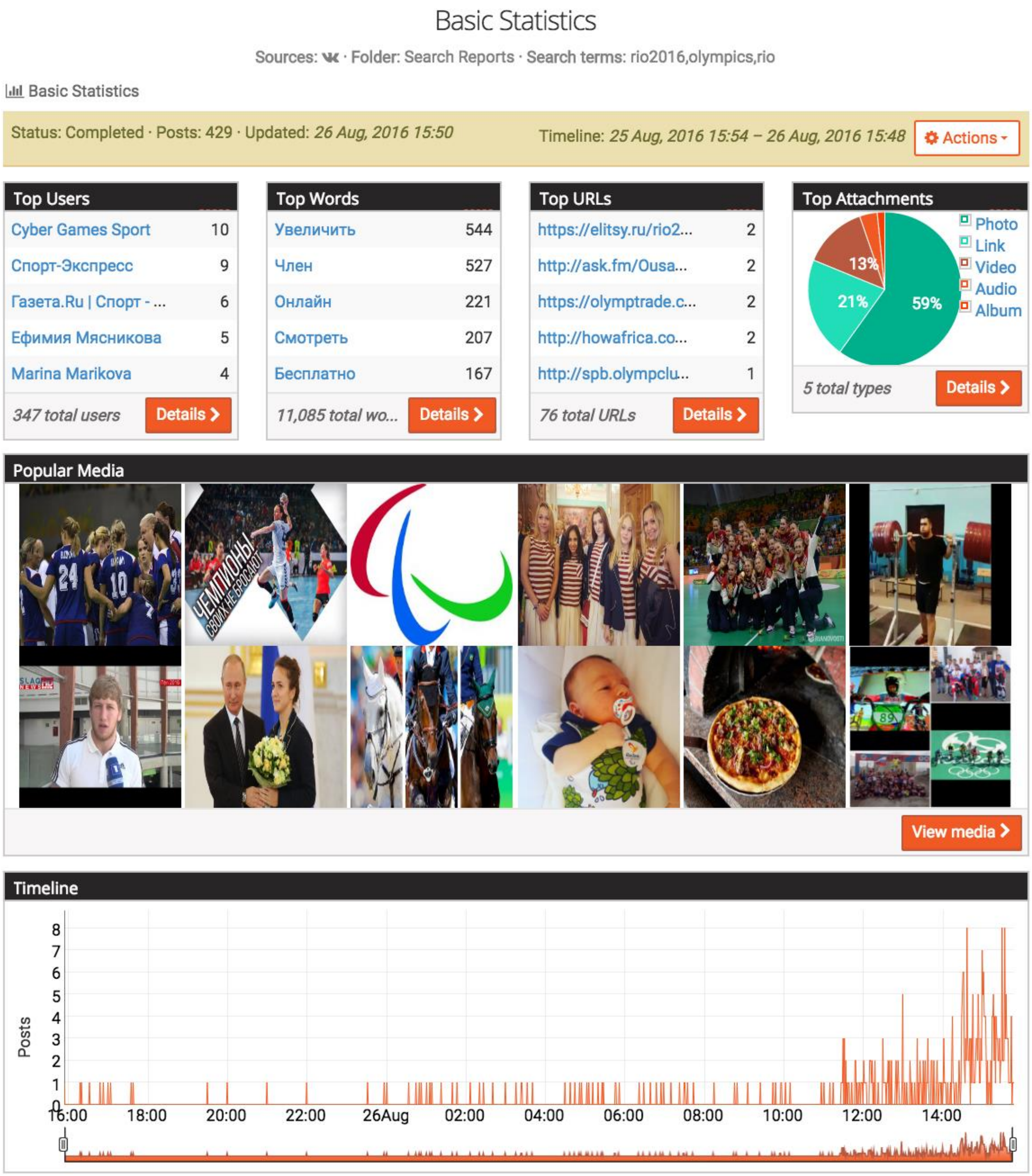
Current limit: 500 – Total quota left in August, 2016: 45,542

Sets the limit for the number of posts in the report

Figure 10: Representative VK Search Screen.

2.5.2 VK Analytics

We have also developed capabilities to compute basic statistics of the VK posts that were collected using the interface of Figure 10. In particular, top users, top words, top URLs, and top attachments along with top popular media is computed and presented to the user in one screen as shown in Figure 11. The timeline of the posts (not shown) is also presented in the same screen. Each top statistics (e.g., top words) can be further drilled down to show additional information.



Basic Statistics

Sources: vk · Folder: Search Reports · Search terms: rio2016,olympics,rio

Basic Statistics

Status: Completed · Posts: 429 · Updated: *26 Aug, 2016 15:50*

Timeline: *25 Aug, 2016 15:54 – 26 Aug, 2016 15:48*

Top Users	
Cyber Games Sport	10
Спорт-Экспресс	9
Газета.Ru | Спорт - ...	6
Ефимия Мясникова	5
Marina Marikova	4
347 total users	

Top Words	
Увеличить	544
Член	527
Онлайн	221
Смотреть	207
Бесплатно	167
11,085 total wo...	

Top URLs	
https://elitsy.ru/rio2...	2
http://ask.fm/Ousa...	2
https://olymptrade.c...	2
http://howafrica.co...	2
http://spb.olympclu...	1
76 total URLs	

Top Attachments

5 total types

Figure 11: Representative VK Basic Statistics View.

Map visualization and NER have also been tailored for VK.

3 Documentation of Features

Please see the Scraawl help page: https://blog.scraawl.com/faq/

4 Conclusion

The proposed work is expected to support a variety of Navy Operational needs in the areas of Disaster Relief Operations (DRO) and assisting decision makers and field personnel to better understand and address unexpected crises. By using graph analytics at large-scales, users will be able to find key actors, extract relationships between these key actors, find pathways between leaders and followers, and interactively visualize social graphs and friendship graphs. In addition, with the added YouTube and VK capabilities, user will be able to get a broader range of textual and relationship data, and will be able to support HSCB modeling, decision making, and COA development.